## **H1B-Visa Approval Prediction using Machine Learning**

# 1 INTRODUCTION

* 1. Overview

Over 2 million visa petitions are filed by the employers each year and only 65000 petitions are approved. So, the goal is to explore the petitions filed and their outcomes for the past six years i.e., from 2011 to 2016, and to find a pattern to predict the outcome by using a predictive model developed using Machine Learning techniques.

In the Guided Project, our goal is to predict the outcome of H-1B visa applications that are filed by many professional foreign nationals every year. Here, we framed the problem as a classification problem and applied it in order to output a predicted case status of the application. The input to our algorithm is the attributes of the applicant. H-1B is a type of non-immigrant visa in the United States that allows foreign nationals to work in occupations that require specialized knowledge and a bachelor’s degree or higher in the specific specialty. This visa requires the applicant to have a job offer from an employer in the US before they can file an application to the US immigration service (USCIS). We believe that this prediction algorithm could be a useful resource both for the future H-1B visa applicants and the employers who are considering sponsoring them.

In order to predict the case status of the applicants, we will be feeding the model with the dataset which contains the required fields by which the machine can classify the case status as certified or denied.

1.2 Purpose

The project's goal is to extract the libraries for machine learning for Visa prediction using Python's pandas, matplotlib, and seaborn libraries.

Next step is to do an exploratory analysis of the dataset to answer questions like: What are the top companies that have applied to the H-1B for employees? What is the trend of the total number of H-1B applications? What is the top popular Job Title and Worksites for H-1B Visa holders?

Third step is to deploy a web application that predicts visa status based on the best performing machine learning algorithms. This feature will help employees to get a real-time prediction based on previous years data.

### By the end of this project:

* You will be able to understand the problem to classify the problem to classify if it is a regression or a classification kind of problem.
* You will be able to know how to pre-process/clean the data using different data pre-processing techniques.
* Applying different algorithms according to the dataset.
* You will be able to know how to find the accuracy of the model.
* You will be able to build web applications using the Flask framework.

# 2 LITERATURE SURVEY

2.1 Existing problem

The practise of evaluating data from many viewpoints and extracting meaningful knowledge from it is known as data mining. It is at the heart of the process of knowledge discovery. Classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression, and other data mining techniques are examples. The most widely used data mining technique is classification, which uses a group of pre-classified samples to create a model that can categorise the entire population of information. The categorization technique is particularly well suited to fraud detection and credit risk applications. This method often employs a classification algorithm based on decision trees. A training set is used to develop a model as a classifier that can categorise data objects into their respective classes in classification. The model is validated using a test set.

2.2 Proposed solution

Our model and analysis will provide a whole picture of the different approval rates by comparing different conditions based on previous data. In addition, our analysis will also provide some statistic data to visualize the characteristics of the application case and trends. In order to predict the status, we will be training the model with occupation category, prevailing wage, Year of application and Job duration after removing the outliers and applying label encoding to all the categorical data. For analysis part we’ll be plotting different graphs to get a relevant inference and eye appealing layout.

# 3 THEORITICAL ANALYSIS

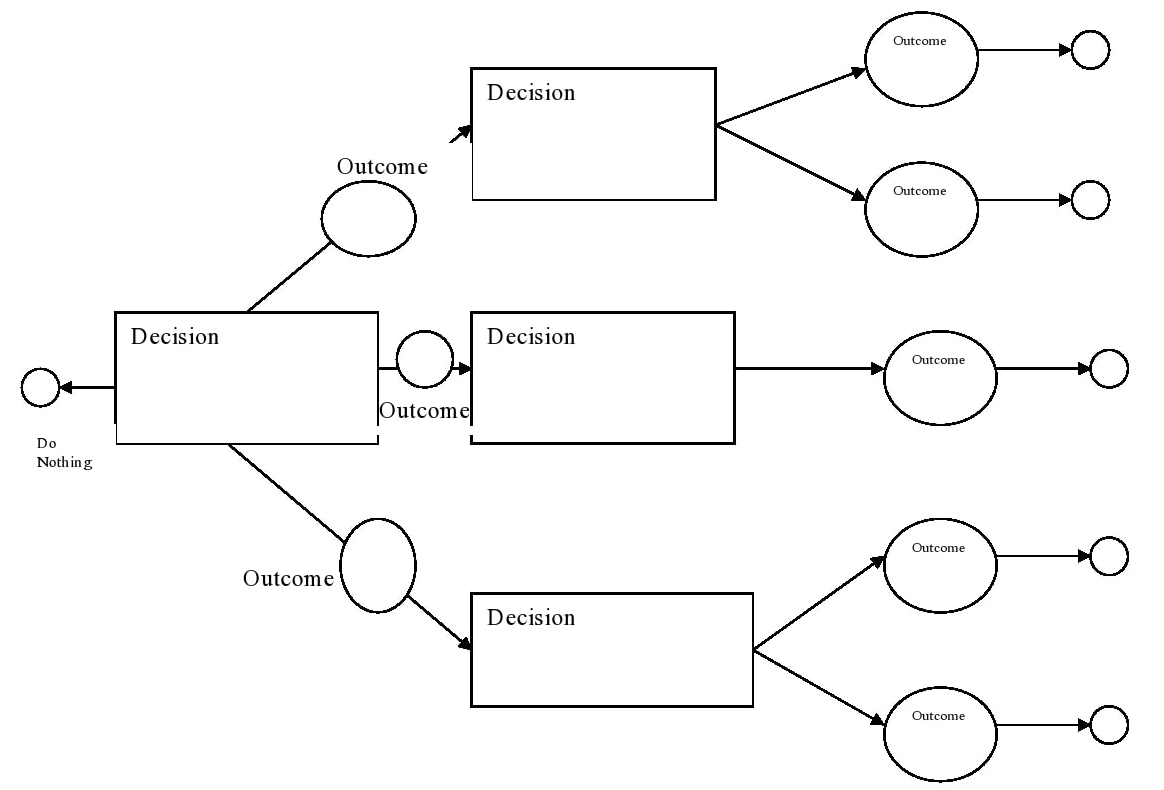
3.1 Block diagram

While selecting the algorithm that gives an accurate prediction, we gone through lot of algorithms like Decision tree, Random Forest etc., which gives the results abruptly accurate and from them we selected only one algorithm for the prediction problem that is Logistic Regression (because it gave a better accuracy). The peculiarity of this problem is collecting the customers details real time and working with the prediction at the same time, so we developed a user interface for the people who'll be accessing for the Visa status prediction.

1. Algorithms Used

1.1 Decision Tree

Decision trees model sequential decision problems under uncertainty. A decision tree describes graphically the decisions to be made, the events that may occur, and the outcomes associated with combinations of decisions and events. Probabilities are assigned to the events, and values are determined for each outcome. A major goal of the analysis is to determine the best decisions. The model of the decision tree was illustrated in the Figure 1.



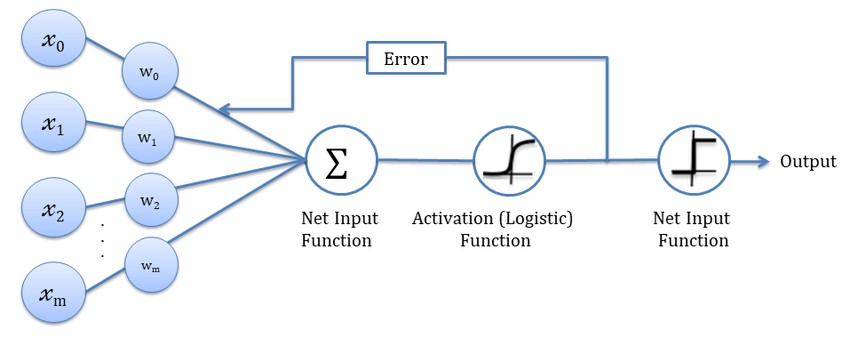
1.2 Logistic regression

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for given set of features (or inputs), X. Contrary to popular belief, logistic regression IS a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself.

Logistic regression used in our project is Multinomial Logistic Regression.

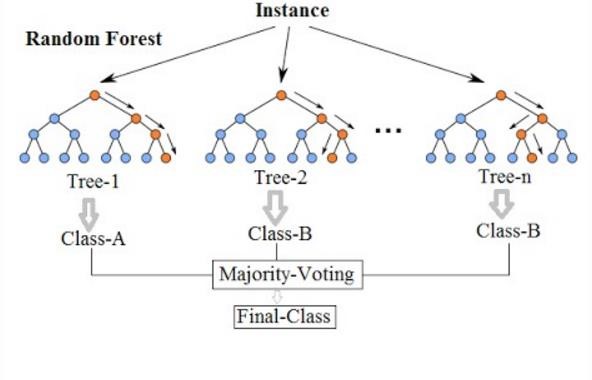
• In Multinomial Logistic Regression, the output variable can have more than two possible discrete outputs. Consider the Digit Dataset. Here, the output variable is the digit value which can take values out of (0, 12, 3, 4, 5, 6, 7, 8, 9).

• It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large sample approximations.



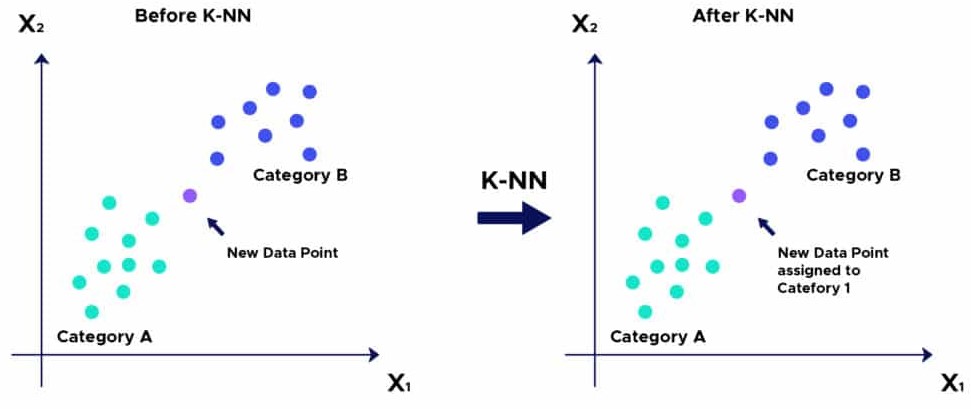
1.3 Random Forest classification

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model’s prediction The fundamental concept behind random forest is a simple but powerful one: A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don’t constantly all err in the same direction).



1.4 K-Nearest Neighbour

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.



3.2 Software designing

We used the Python programming language, which is an interpreted and high-level programming language, and Machine Learning techniques to create this Visa Approval status forecast. For coding, we used the Anaconda distribution's Jupyter Notebook environment and the Spyder, which is an integrated scientific programming language in Python. We utilised Flask to create a user interface for the prediction. It's a Python-based microweb framework. Because it does not require any specific tools or libraries, it is considered as a micro framework. It has no database abstraction layer, form validation, or any other components where a

pre-existing third-party libraries provide common tasks, and the only scripting language for building a webpage is HTML, which is done by creating templates to use in the Flask and HTML functions.

**4 EXPERIMENTAL INVESTIGATIONS**

In this paper, the dataset we used is derived from H-1B\_Kaggle. It contains more than 30L H-1B Visa data of users.

1. CASE STATUS: Status associated with the last significant event or decision. We labelled ’CERTIFIED’ case as 0 ‘CERTIFIED-WITHDRAWN’ case as 1 ‘DENIED’ case as 2 ’WITHDRAWN’ case as 3 ’PENDING QUALITY AND COMPILANCE-UNASSIGNED’ case as 4 ‘REJECTED’ case as 5 and ‘INVALIDATED’ cases as 6.

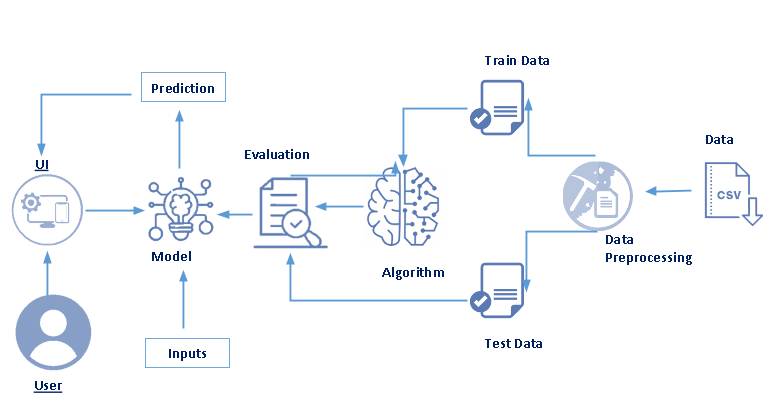
2. EMPLOYER\_NAME: Name of employer submitting the H1-B application. Used in comparing salaries and number of applications of various employers.

3. JOB\_TITLE: Title of the job using which we can filter specific job positions for e.g., Data Scientist, Data Engineer etc.

4. PREVAILING\_WAGE: The average wage paid to similarly employed workers in the sought occupation in the planned employment location is referred to as the prevailing wage for a job position. The prevailing wage is determined by the employer's basic job requirements. (Source). Tis column will be One of the primary measures of the data analysis.

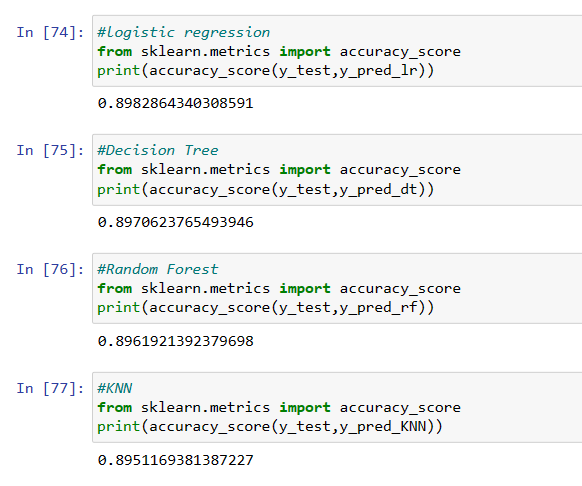
5. WORKSITE\_CITY, WORKSITE Site: The intended field of employment for the foreign worker. We'll look at the relationship between the prevailing wage for Data Scientist jobs in various places.

**5 FLOW CHART**



**6 RESULT**

In this project, we have used logistic regression for the prediction and compared it with other algorithms like random forest, decision tree and KNN but between all logistic regression gives best prediction percentage of 89.82.

****

**7 ADVANTAGES & DISADVANTAGES**

**Advantages**

* Logistic regression is one of the most basic machine learning algorithms. It is simple to build and, in some situations, delivers excellent training efficiency. Because of these factors, training a model with this technique does not necessitate a lot of computing resources.
* The inferences regarding the importance of each characteristic are based on the expected parameters (trained weights). The association's orientation, positive or negative, is also specified. As a result, logistic regression can be used to determine the relationship between the features.
* Along with classification findings, Logistic Regression produces well calibrated probability. This is a benefit over models that just provide results for the final classification.

**Disadvantages**

* Logistic regression is a statistical analysis technique that uses independent features to try to predict precise probability outcomes. On high-dimensional datasets, this may cause the model to be over-fit on the training set, overstating the accuracy of predictions on the training set, and so preventing the model from accurately predicting results on the test set. This is most common when the model is trained on a little amount of training data with a large number of features. Regularization strategies should be considered on high-dimensional datasets to minimise over-fitting (but this makes the model complex).
* The model may be under-fit on the training data if the regularisation variables are too high. Because logistic regression has a linear decision surface, it cannot tackle nonlinear issues. In real-world circumstances, linearly separable data is uncommon. As a result, non-linear features must be transformed, which can be done by increasing the number of features such that the data can be separated linearly in higher dimensions.

**8 APPLICATIONS**

The web app made by using ML models could be used by students or employee to check whether their application will be certified or not based on previous years data.

**9 CONCLUSION**

Several machine learning models, such as Random Forest and Naïve Bayes, can be used to predict the outcome of H-1B visa applications based on the applicant's qualities. We tried to use the logistic regression model for the project as it was more convenient for the dataset and it gave a better accuracy. Finally, it's possible to include this into a web application and find out the predictability of the visa.

**10 FUTURE SCOPE**

Further Logistic regression can be applied on other data sets available for visa approvals to further investigate its accuracy. Other machine learning algorithms can also be implemented in the project like Naïve bayes model or the SVM model.In further study, we will try to conduct experiments on larger data sets or try to tune the model so as to achieve the state -of-art performance of the model and a great UI support system making it complete web application model. The project can also probe deeper in the process of predicting visa for an individual by including Job title, location and also through categorisation of the individuals.

**11 BIBILOGRAPHY**

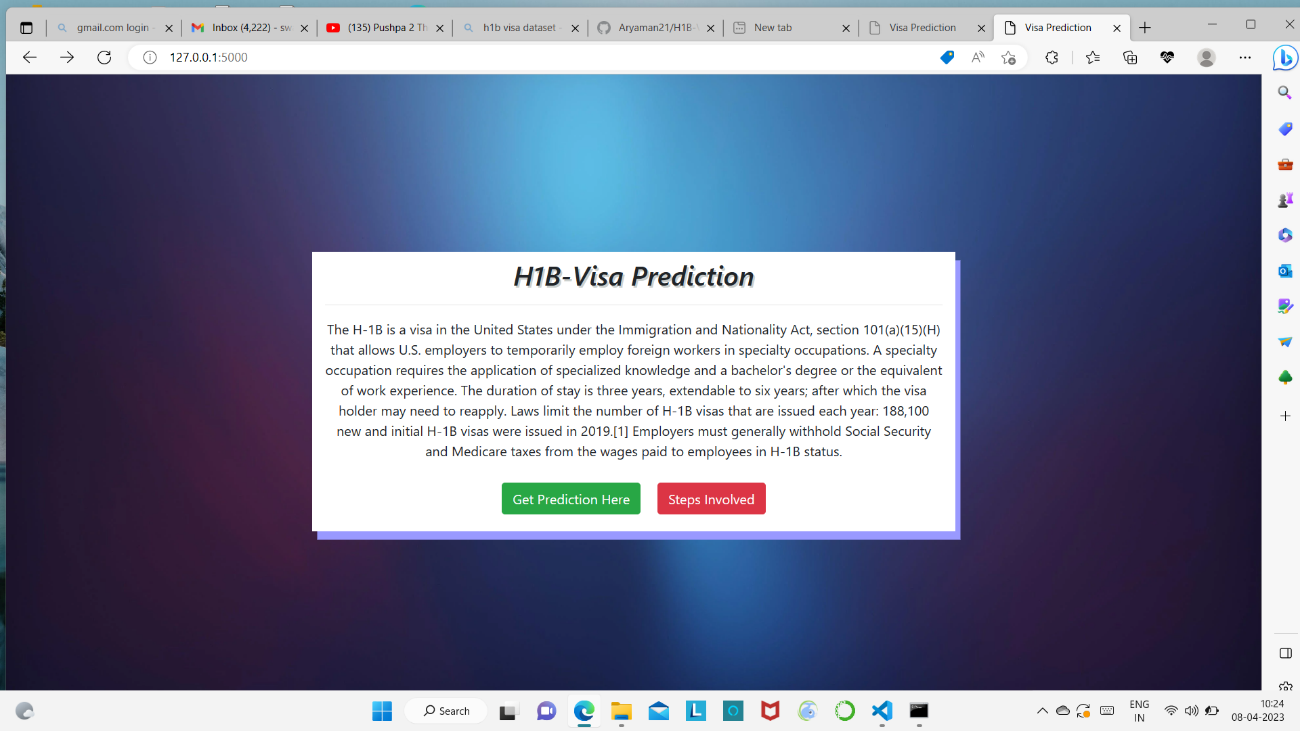
<https://www.kaggle.com/nsharan/h-1b-visa><https://www.immi-usa.com/h1b-visa/h1b-visa-benefits/><https://www.javatpoint.com/logistic-regression-in-machine-learning>[https://iq.opengenus.org/advantages-and-disadvantages-of-logisticregression/](https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/) [https://towardsdatascience.com/optimizing-hyperparameters-in-random-forestclassification-ec7741f9d3f6](https://towardsdatascience.com/optimizing-hyperparameters-in-random-forest-classification-ec7741f9d3f6) [https://towardsdatascience.com/understanding-random-forest-](https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree)

[58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20 any%20individual%20tree](https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree).

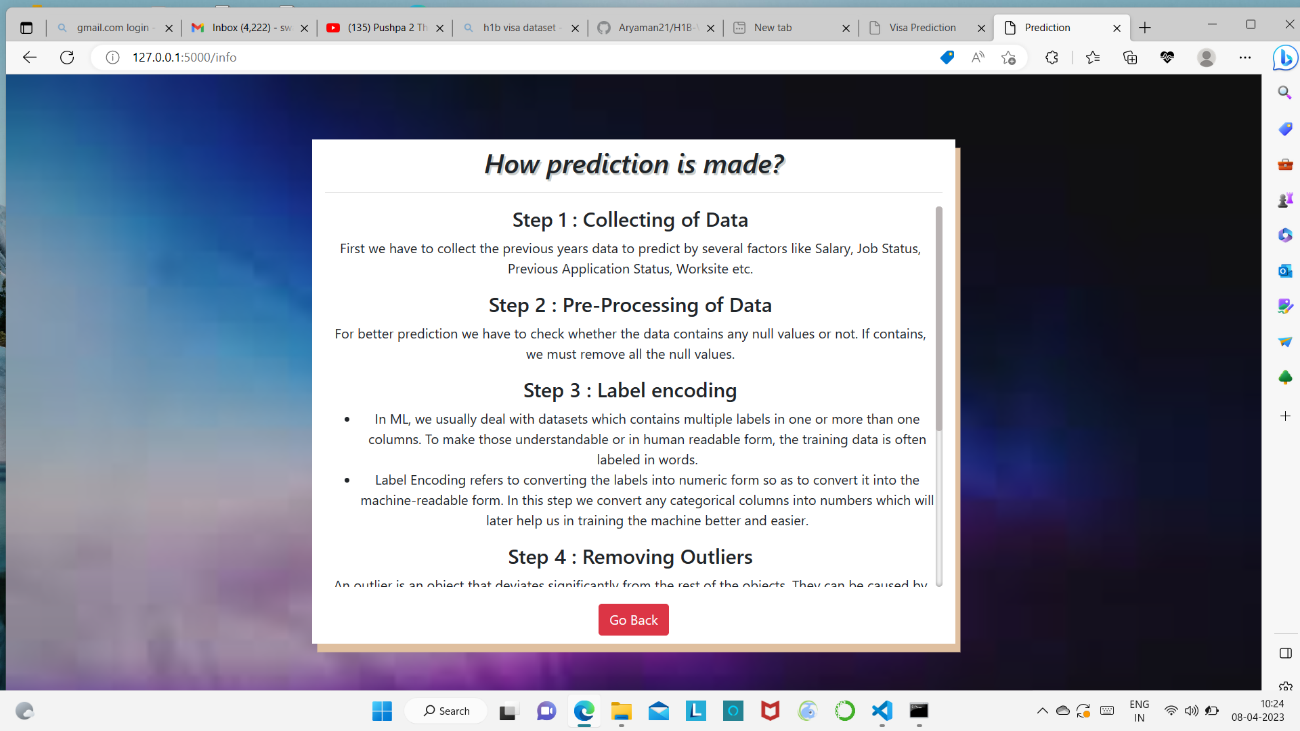
**APPENDIX**

SCREENSHOTS OF PREDICTION WEBSITE

HOME PAGE



INFO PAGE



INDEX PAGE

